Decision Theory for Statistical Inference

Decision Rules, Loss, Risk, Admissibility, Bayes and Minimax Procedures

Lecture notes, Statistical Inference Session 3

1 Introduction

Netflix is comparing its current recommendation algorithm, version A, with a proposed algorithm, version B. Users are randomly assigned to one of the two versions. First, Netflix identifies the five shows with the highest total platform-wide viewing minutes during the previous week. For each experimental user, the outcome is the number of minutes that user watches in those five shows. Thus, the set of five shows is defined by prior-week viewing on the platform, not by where the recommendation algorithm places them. Let

$$\theta = \mu_B - \mu_A$$

denote the true average change in this metric caused by deploying B rather than A. A positive θ means that B increases viewing in the five shows identified as the platform's most watched during the previous week.

An experiment produces an estimate X of θ; in a large experiment it is reasonable to begin with the approximation

$$X \mid \theta \sim N(\theta, \sigma^2),$$

where σ^2 is the sampling variance of the estimated difference. Reporting X is an estimation problem. Choosing whether to deploy B is a decision problem. The distinction matters: an observed gain of one minute may be statistically informative but too small to justify engineering, product, or user-experience costs. Conversely, failing to deploy a genuinely better recommender has an opportunity cost.

Decision theory separates four ingredients:

$$\text{model} \quad + \quad \text{available actions} \quad + \quad \text{consequences of actions} \quad + \quad \text{decision rule.}$$

This formulation forces us to state what a good procedure means. In the Netflix example, "good" might mean accurately estimating the lift, maximizing expected viewing after accounting for rollout cost, or guarding against a disastrous reduction in viewing. There is usually no uniformly best rule without a criterion for comparing its performance across parameter values.

2 The anatomy of a statistical decision problem

Let X denote the observed data, taking values in a sample space $\mathcal{X}$. Its distribution belongs to a family $\{P_\theta : \theta \in \Theta\}$.

Running example. For the Netflix experiment, the state is the unknown lift θ. If the goal is estimation, the action space is $\mathcal{A} = \mathbb{R}$ and an action is a reported estimate of the change in minutes watched. If the goal is rollout, the action space is $\mathcal{A} = \{a_A, a_B\}$, where a_A means retain algorithm A and a_B means deploy algorithm B.

Definition 2.1 (Action space)**.** *The action space $\mathcal{A}$ is the set of actions available after observing the data. In point estimation one often has $\mathcal{A} = \Theta$; in a two-action problem, $\mathcal{A} = \{a_0, a_1\}$.*

Definition 2.2 (Decision rule)**.** *A nonrandomized decision rule is a measurable function*

$$\delta : \mathcal{X} \longrightarrow \mathcal{A}.$$

After observing $X = x$, we take action $\delta(x)$. A randomized rule specifies a probability distribution on $\mathcal{A}$ for each x.

An estimator is therefore a special kind of decision rule. If the action is a reported value for θ, then $\delta(X) = \widehat{\theta}$. The term *decision function* is often used synonymously with decision rule.

Definition 2.3 (Loss)**.** *A loss function is a map*

$$L : \Theta \times \mathcal{A} \longrightarrow [0, \infty),$$

where $L(\theta, a)$ quantifies the consequence of taking action a when the true state is θ.

Common estimation losses include

$$\begin{aligned}
&L(\theta, a) = (a - \theta)^2 && \text{(squared-error loss)},\\
&L(\theta, a) = |a - \theta| && \text{(absolute-error loss)},\\
&L(\theta, a) = \frac{(a - \theta)^2}{\theta^2}, \quad \theta \neq 0 && \text{(relative squared-error loss)},\\
&L(\theta, a) = \mathbf{1}\{a \neq \theta\} && \text{(zero-one loss, for discrete actions)}.
\end{aligned}$$

The choice is substantive, not merely mathematical. Squared loss heavily penalizes large errors, whereas absolute loss is less sensitive to them.

For Netflix, squared-error loss may be used to judge an estimate of the lift:

$$L(\theta, a) = (a - \theta)^2.$$

For the rollout decision, the losses need not be symmetric. Deploying B when it truly reduces viewing can impose product and rollback costs, while retaining A when B would increase viewing sacrifices engagement. A stylized loss is

$$L(\theta, a_B) = C_D \mathbf{1}\{\theta \leq 0\}, \qquad L(\theta, a_A) = C_M \theta \mathbf{1}\{\theta > 0\},$$

where C_D is the cost of a harmful deployment and $C_M\theta$ represents the opportunity cost of missing a positive lift. Real decisions could also include latency, satisfaction, retention, and content-diversity costs; the watched-minutes metric alone does not determine the loss function.

Definition 2.4 (Risk)**.** *The frequentist risk function of a rule δ is its expected loss under repeated sampling at state θ:*

$$R(\theta, \delta) = \mathbb{E}_\theta[L\{\theta, \delta(X)\}].$$

Thus, $R(\cdot, \delta)$ is a function on Θ, not generally a single number.

For a rollout rule δ, $R(\theta, \delta)$ answers a concrete repeated-experiment question: if the true lift were θ and Netflix repeatedly ran experiments of this size, what average loss would result from applying the same rollout rule each time? Sampling variation makes the rule sometimes deploy B and sometimes retain A even though the underlying θ is fixed.

The expectation in the risk is over the sampling distribution of X while θ remains fixed. This is distinct from a Bayes calculation, in which θ is averaged with respect to a prior.

2.1 Bias–variance decomposition

Under squared-error loss, the risk of an estimator $\delta(X)$ equals its mean squared error:

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[(\delta - \theta)^2] \\ &= \mathrm{Var}_\theta(\delta) + \{\mathbb{E}_\theta(\delta) - \theta\}^2. \end{aligned}$$

The first term is variance and the second is squared bias. Decision theory therefore makes clear that unbiasedness is not by itself an optimality criterion: a small bias may purchase a large reduction in variance.

Example 2.5 (Normal mean and shrinkage toward zero)**.** *Let $X \sim N(\theta, 1)$ and consider $\delta_c(X) = cX$, where $0 \le c \le 1$. Under squared-error loss,*

$$R(\theta, \delta_c) = \mathbb{E}_\theta[(cX - \theta)^2] = c^2 + (1 - c)^2\theta^2.$$

For the usual estimator $\delta_1 = X$, $R(\theta, \delta_1) = 1$. For a fixed $c < 1$, shrinkage improves risk near zero but becomes worse for sufficiently large $|\theta|$. Indeed,

$$R(\theta, \delta_c) < 1 \iff |\theta| < \sqrt{\frac{1+c}{1-c}}.$$

This crossing of risk curves is why a further comparison criterion is needed.

In the Netflix setting, X is the unshrunk estimated lift and cX shrinks it toward zero, expressing skepticism that a new algorithm creates a large effect. Shrinkage reduces noisy overstatement when the true lift is near zero, but it can understate a genuinely large improvement. The risk calculation makes that tradeoff explicit.

3 Dominance and admissibility

Definition 3.1 (Dominance)**.** *A rule δ_1 dominates δ_2 if*

$$R(\theta, \delta_1) \le R(\theta, \delta_2) \quad \textit{for every } \theta \in \Theta,$$

with strict inequality for at least one θ.

Definition 3.2 (Admissibility)**.** *A rule is admissible if no other rule dominates it. Otherwise it is inadmissible.*

Admissibility is a minimal rationality requirement: an inadmissible rule can be replaced without increasing risk anywhere and with a strict improvement somewhere. It does *not* say that an admissible rule is uniquely best. Many rules may be admissible, and their risks may cross.

For Netflix, a rollout policy is inadmissible if another policy has no greater expected loss for every possible true lift and lower loss for at least one lift. Such a policy should be discarded before arguing about priors or worst-case protection. Admissibility therefore acts as a screening principle, not as a complete prescription for deployment.

Example 3.3 (A visibly inadmissible binomial estimator)**.** *Let $X \sim \mathrm{Bin}(n, p)$ and estimate $p \in [0, 1]$ under squared-error loss. Compare*

$$\delta_0(X) = 0 \quad \textit{and} \quad \delta_1(X) = \frac{X}{n}.$$

Their risks are

$$R(p, \delta_0) = p^2, \qquad R(p, \delta_1) = \frac{p(1-p)}{n}.$$

Neither dominates the other: δ_0 is excellent at $p = 0$ and poor away from zero. Now consider the artificial rule $\delta_2(X) = 2$. Since $p \in [0, 1]$, projecting any estimate onto $[0, 1]$ cannot increase squared loss. In particular, the constant rule 1 satisfies

$$(1-p)^2 < (2-p)^2 \qquad \textit{for every } p \in [0, 1].$$

Thus δ_2 is dominated and inadmissible. More generally, under squared loss, an estimator taking values outside a convex parameter space can often be improved by projection.

Remark 3.4 (Dimension matters)**.** *For estimating a multivariate normal mean under total squared-error loss, the usual estimator X is admissible in dimensions one and two but inadmissible in dimensions three and above. The James–Stein phenomenon is a famous warning that seemingly obvious procedures can fail admissibility in higher dimensions.*

4 Bayes rules

Frequentist risk leaves one value for each θ. A Bayes criterion combines these values using a prior distribution π on Θ.

Definition 4.1 (Bayes risk and Bayes rule)**.** *The Bayes risk of δ under prior π is*

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\, \pi(d\theta).$$

A Bayes rule δ_π minimizes $r(\pi, \delta)$ over the permitted class of rules.

By iterated expectation,

$$r(\pi, \delta) = \int_{\mathcal{X}} \int_\Theta L\{\theta, \delta(x)\}\, \pi(d\theta \mid x)\, m_\pi(x)\, dx,$$

where m_π is the prior predictive distribution. Hence a Bayes rule may be found pointwise:

$$\delta_\pi(x) \in \operatorname*{arg\,min}_{a \in \mathcal{A}} \mathbb{E}_\pi[L(\theta, a) \mid X = x].$$

The quantity being minimized is the *posterior expected loss.*

In the Netflix experiment, a prior for θ can be informed by earlier recommendation experiments. If historical changes usually produce lifts close to zero, a prior centered near zero is plausible. The posterior combines that history with the current A/B test. A Bayes rollout rule then chooses the algorithm with smaller posterior expected business loss, rather than deploying merely because the observed difference is positive.

Proposition 4.2 (Bayes actions under common losses)**.** *For a fixed observation x:*

1. *Under squared-error loss, a Bayes action is the posterior mean* $\mathbb{E}(\theta \mid x)$.
2. *Under absolute-error loss, a Bayes action is any posterior median.*
3. *Under zero-one loss on a discrete parameter space, a Bayes action is a posterior mode.*

Proof. For squared loss, write $m(x) = \mathbb{E}(\theta \mid x)$. Then

$$\mathbb{E}[(a-\theta)^2 \mid x] = (a - m(x))^2 + \mathrm{Var}(\theta \mid x),$$

so $a = m(x)$ minimizes posterior expected loss. For absolute loss, a subgradient in a is $P(\theta < a \mid x) - P(\theta > a \mid x)$; zero belongs to the subgradient exactly at a posterior median. Under zero-one loss, posterior expected loss is $1 - P(\theta = a \mid x)$, minimized by maximizing posterior probability. □

Example 4.3 (Normal–normal Bayes estimator)**.** *Let*

$$X \mid \theta \sim N(\theta, \sigma^2), \qquad \theta \sim N(\mu_0, \tau^2).$$

The posterior distribution is normal with mean

$$\mu_1 = \frac{\tau^2}{\tau^2 + \sigma^2} X + \frac{\sigma^2}{\tau^2 + \sigma^2} \mu_0$$

and variance

$$v_1 = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}.$$

Under squared-error loss, $\delta_\pi(X) = \mu_1$*. It is a weighted average of the data and the prior mean. If* $\mu_0 = 0$*, it is the shrinkage rule* cX *with* $c = \tau^2/(\tau^2 + \sigma^2)$.

Its Bayes risk is especially simple. Since the posterior variance is constant,

$$r(\pi, \delta_\pi) = \mathbb{E}[v_1] = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}.$$

Interpreted for Netflix, take X to be the estimated difference in watched minutes and μ_0 to be the mean lift across comparable past experiments. The Bayes estimate shrinks the current experimental estimate toward the historical mean. The amount of shrinkage increases when the current experiment is noisy (large σ^2) and decreases when historical effects are heterogeneous (large τ^2).

Example 4.4 (Beta–binomial Bayes estimator)**.** *Let* $X \mid p \sim \mathrm{Bin}(n, p)$ *and* $p \sim \mathrm{Beta}(\alpha, \beta)$*. The posterior is*

$$p \mid X = x \sim \mathrm{Beta}(\alpha + x, \beta + n - x).$$

Under squared-error loss,

$$\delta_\pi(x) = \mathbb{E}[p \mid x] = \frac{\alpha + x}{\alpha + \beta + n} = \frac{n}{n + \alpha + \beta} \frac{x}{n} + \frac{\alpha + \beta}{n + \alpha + \beta} \frac{\alpha}{\alpha + \beta}.$$

The estimator shrinks the sample proportion toward the prior mean.

4.1 When does Bayes imply admissible?

Proposition 4.5 (A useful sufficient condition)**.** *Suppose δ_π is a Bayes rule with finite Bayes risk. If it is the unique Bayes rule up to prior-predictive null sets and the prior gives positive mass to every open subset of Θ, then, under standard regularity conditions, δ_π is admissible.*

The essential argument is short. If another rule dominated δ_π, integrating the risk inequality against π would produce no larger Bayes risk. If the improvement occurs on a set to which the prior assigns positive mass, the Bayes risk would be strictly smaller, contradicting Bayes optimality. The qualifications matter: a Bayes rule need not be admissible when the prior ignores parameter regions or when nonuniqueness and null sets intervene.

5 Minimax rules

Bayes analysis averages risk. Minimax analysis protects against the worst parameter value.

Definition 5.1 (Maximum risk and minimax rule)**.** *The maximum risk of δ is*

$$\overline{R}(\delta) = \sup_{\theta \in \Theta} R(\theta, \delta).$$

A minimax rule δ^* satisfies

$$\sup_\theta R(\theta, \delta^*) = \inf_\delta \sup_\theta R(\theta, \delta).$$

Minimax is conservative: it chooses the rule whose worst-case expected loss is smallest. This does not mean the rule minimizes risk at every θ.

This criterion is relevant when Netflix is unwilling to assign a credible prior to the lift but wants protection across a stated range of plausible effects. A minimax rollout rule is chosen for its worst-case performance. That protection can be valuable for a high-stakes global rollout, but it may be overly conservative when reliable evidence from earlier experiments is available.

Example 5.2 (Normal mean)**.** *Let $X \sim N(\theta, \sigma^2)$ with $\Theta = \mathbb{R}$, and use squared-error loss. The estimator $\delta_0(X) = X$ has constant risk*

$$R(\theta, \delta_0) = \sigma^2.$$

To establish minimaxity, consider proper priors $\pi_\tau = N(0, \tau^2)$. From the normal–normal calculation, their minimum Bayes risks are

$$r(\pi_\tau, \delta_{\pi_\tau}) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \longrightarrow \sigma^2 \qquad (\tau^2 \to \infty).$$

For every rule δ and every prior π,

$$\sup_\theta R(\theta, \delta) \geq \int R(\theta, \delta)\, \pi(d\theta) \geq \inf_{\widetilde{\delta}} r(\pi, \widetilde{\delta}).$$

Therefore every rule has maximum risk at least σ^2 in the limit, while X attains maximum risk σ^2. Hence X is minimax.

Proposition 5.3 (Constant-risk Bayes rule)**.** *If a Bayes rule δ_π has constant risk $R(\theta, \delta_\pi) = c$ for all θ, then it is minimax, provided its Bayes risk equals c.*

Proof. For any rule δ,

$$\sup_\theta R(\theta, \delta) \geq r(\pi, \delta) \geq r(\pi, \delta_\pi) = c.$$

But $\sup_\theta R(\theta, \delta_\pi) = c$, so c is the minimax value. □

Definition 5.4 (Least favorable prior)**.** *A prior is least favorable if its minimum Bayes risk is at least that of every other prior. It places weight on parameter values that make the decision problem hardest.*

When a least favorable prior exists and its Bayes rule has maximum risk equal to its Bayes risk, that Bayes rule is minimax. Least favorable priors provide an important bridge between Bayes and minimax reasoning.

6 A two-state decision problem and the two kinds of error

We now simplify the Netflix example to a binary state space without developing hypothesis-testing machinery. Let state 0 mean that algorithm B provides no worthwhile improvement over A, and let state 1 mean that B provides a worthwhile improvement. The actions are a_A (retain A) and a_B (deploy B).

Use the loss table

	a_A	a_B
state 0	0	C_{10}
state 1	C_{01}	0

where C_{10} is the cost of deploying B when it is not worthwhile, and C_{01} is the opportunity cost of retaining A when B is worthwhile.

For a decision rule $\delta(X) \in \{a_A, a_B\}$, define

$$\alpha = P_0\{\delta(X) = a_B\}, \qquad \text{Type I error: deploy B when it is not worthwhile,}$$
$$\beta = P_1\{\delta(X) = a_A\}, \qquad \text{Type II error: retain A when B is worthwhile.}$$

These names describe the direction of the mistake; they do not by themselves say which mistake is more serious. For Netflix, a Type I error may cause a costly rollout or reduce user experience; a Type II error leaves potential viewing gains unrealized. The risk function is simply

$$R(0, \delta) = C_{10}\alpha, \qquad R(1, \delta) = C_{01}\beta.$$

Thus Type I and Type II errors are not ideas separate from decision theory: they are the two state-specific components of risk in a binary decision problem.

6.1 Bayes rule for two simple states

Let $f_0(x)$ and $f_1(x)$ denote the sampling densities under states 0 and 1, and let $P(\theta = 1) = \pi_1$, $P(\theta = 0) = \pi_0$. Given $X = x$, the posterior expected losses are

$$\rho(a_B \mid x) = C_{10}P(\text{state } 0 \mid x),$$
$$\rho(a_A \mid x) = C_{01}P(\text{state } 1 \mid x).$$

Choose a_B exactly when

$$\frac{f_1(x)}{f_0(x)} > \frac{C_{10}\pi_0}{C_{01}\pi_1}.$$

The threshold reflects both prior odds and relative consequences. Equal error costs and equal prior probabilities give a likelihood-ratio threshold of 1.

Example 6.1 (One normal observation)**.** *Let X denote a standardized difference in watched minutes and suppose*

$$X \mid \theta = 0 \sim N(0,1), \qquad X \mid \theta = 1 \sim N(\mu, 1), \quad \mu > 0.$$

For a threshold rule $\delta_c(x) = a_B$ when $x > c$,

$$\alpha(c) = 1 - \Phi(c), \qquad \beta(c) = \Phi(c - \mu).$$

Lowering c decreases Type II error but increases Type I error. No threshold reduces both simultaneously.

The likelihood ratio is

$$\frac{f_1(x)}{f_0(x)} = \exp\left(\mu x - \frac{\mu^2}{2}\right).$$

Hence the Bayes threshold is

$$c_B = \frac{\mu}{2} + \frac{1}{\mu}\log\left(\frac{C_{10}\pi_0}{C_{01}\pi_1}\right).$$

If missing a worthwhile recommendation improvement is more costly, then C_{01} rises and c_B falls: Netflix deploys B more readily, accepting more Type I errors to avoid Type II errors.

6.2 A minimax threshold in the symmetric normal problem

In the preceding example, take $C_{10} = C_{01} = 1$. The maximum risk of threshold c is

$$\max\{1 - \Phi(c), \Phi(c - \mu)\}.$$

The first term decreases with c and the second increases. Their maximum is minimized where they are equal:

$$1 - \Phi(c) = \Phi(c - \mu).$$

Using $1 - \Phi(c) = \Phi(-c)$ and strict monotonicity of Φ, we obtain $-c = c - \mu$, so

$$c_M = \frac{\mu}{2}.$$

At this threshold, the two state-specific risks are equal. It is also the Bayes rule for equal prior probabilities. Here the equal prior is least favorable.